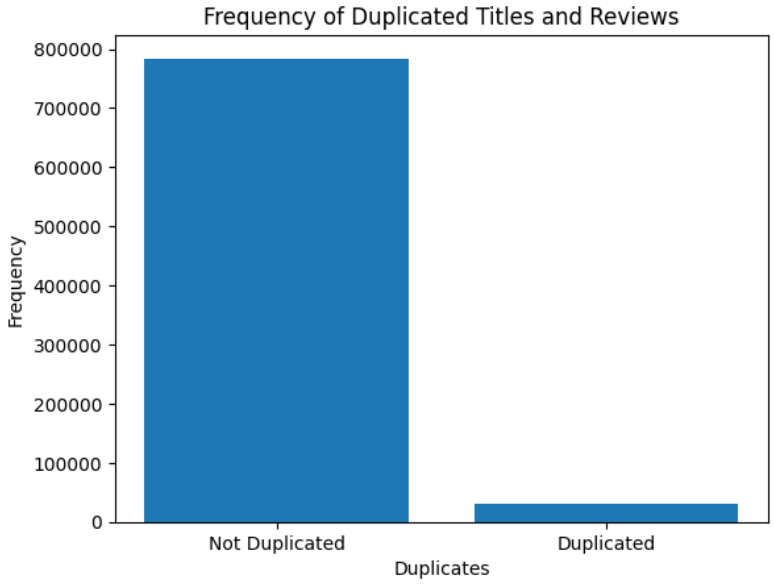
1. **Problem:**
   The primary objective of this project is to perform sentiment analysis on Vietnamese hotel reviews obtained from Booking.com. The aim is to predict the sentiment or score associated with the reviews, reflecting customer feedback.

2. **Data processing approach**
   a. **Cleaning data steps:**
   - Step 1: Remove all NaN (813676 rows left)

	score	title	review
3000	5.0	Friendly and helpful team of staff	NaN
3001	5.0	Friendly and helpful team of staff	NaN
3002	5.0	Friendly and helpful team of staff	NaN
3003	5.0	Friendly and helpful team of staff	NaN
3004	5.0	Friendly and helpful team of staff	NaN

   - Step 2: Remove all duplicate contents (784031 rows left)



   - Step 3: Remove meaning less contents

	score	title	review
268588	3.0	さりげに社会主義のかおりが。。。	...
461960	4.0	オーシャンフロントがオススメ！だけど、蚊多い..	...
582890	5.0	アンビン島でホームステイ	...
698238	5.0	日系ホテルらしいおもてなしが最高！	...
811541	4.0	バルコニーつきデラックスの部屋はイイ感じ！	...

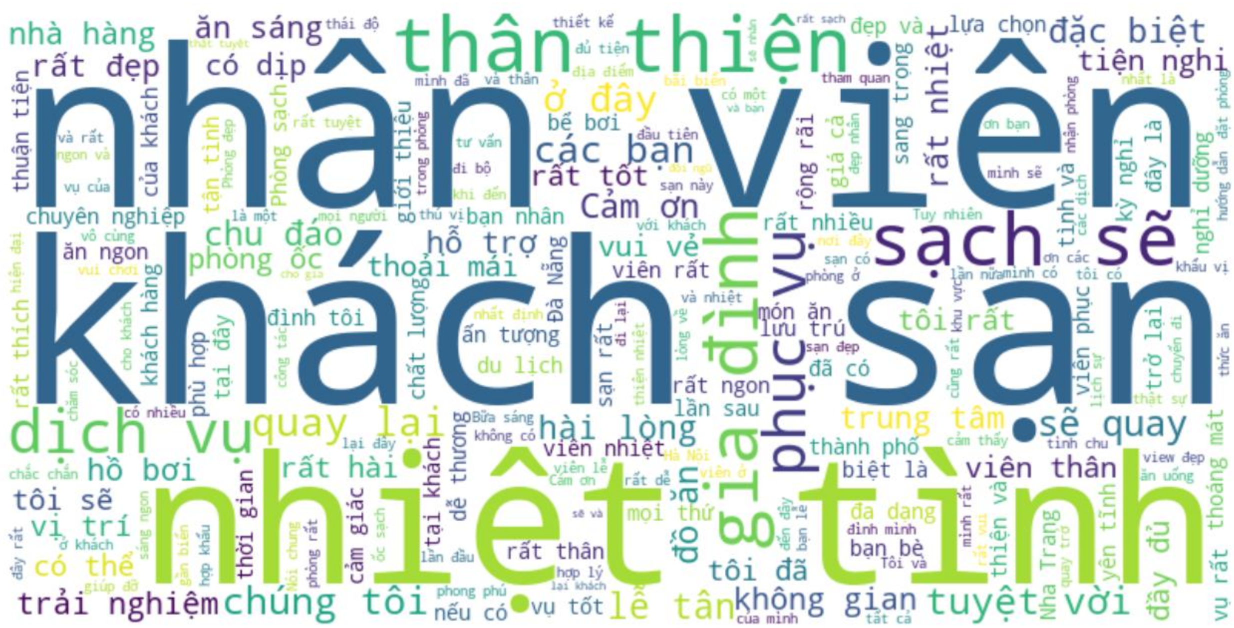
- Step 4: Classify languages for all review using **xlm-roberta-base-language-detection**

	score	title	review	language
0	5.0	Very good hotel	Good hotel i have ever stayed in Vietnam, good...	en
1	4.0	BUEN ALOJAMIENTO QUE GANARIA MUCHO MEJORANDO E...	Este hotel está muy cerca del barrio de las em...	es
2	5.0	Great place in Cau Giay	This place was very nice. Our bedroom were cle...	en
3	5.0	TRẢI NGHIỆM TỐT	Đầy đủ dịch vụ tiện nghi Ăn sáng buffee ngon H...	vi
4	5.0	Perfect stay	It was a amazing hotel. They helped very good ...	en
...	...	...	...	...
784026	5.0	乾淨整潔，交通方便	位於峴港市區，距離韓江橋或韓市場都不會太遠，店員很熱心，還可以幫忙預訂摩托車跟行程，非常值得...	zh
784027	5.0	Check this place	My friend and I received excellent and profess...	en
784028	5.0	店员给了我们很多帮助，装修简单精致，卫生很好	这是我们此行到越南第一个入住的酒店，也是呆的时间最长的酒店。酒店原本是一家咖啡店，其次楼上有...	zh
784029	5.0	Công tác	Rất tuyệt vời... khi đến đây tôi cảm giác thoả...	vi
784030	3.0	Neu, hübsch eingerichtet, aber einige Mängel	Das Hostel ist ganz neu und sehr nett eingeric...	de

b. Exploratory Data Analysis (EDA)

- Vietnamese reviews: 48102 rows

- Apply WordCloud to visualize the most frequently appearing words in the DataFrame. As the dataset pertains to hotels, it's evident that common phrases such as "nhân viên" (staff) and "khách sạn" (hotel) are prevalent.

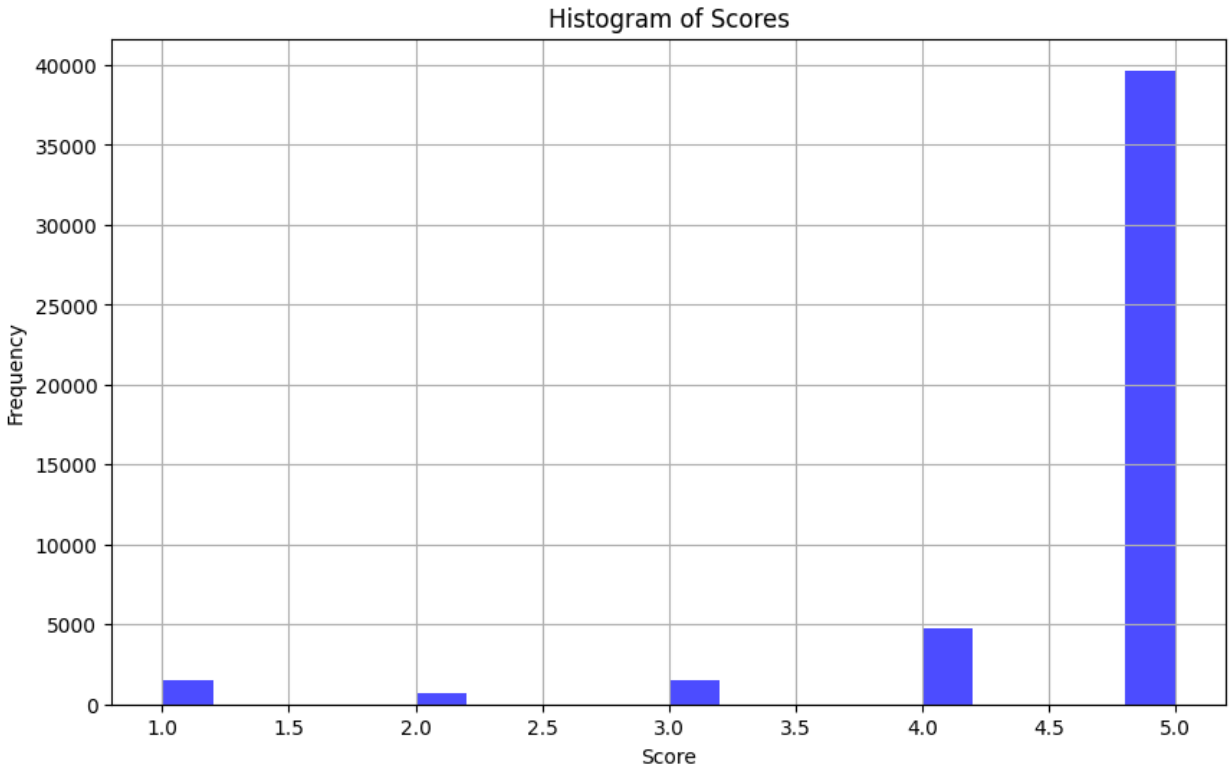


- There is five label in 'score' column, and I saw that some value are Float and some are String, so we need to convert it all to Float type.

- I notice that there are some emoji inside the dataset, and we will also want to predict the reviews contain emoji as well, and most of the emojis are positive.

	emoji	count
0	❤	2260
1	🥰	922
2	😍	873
3	👍	819
4	😘	400
5	😊	334
6	▤	307
7	♥	261
8	😁	163
9	😒	151
10	💕	151
11	😂	150
12	⭐	131
13	👌	108
14	💯	92
15	🤩	85
16	😄	75
17	🤣	66
18	😀	64
19	🌟	64

- Seem like the diversity of each score is really big and unbalance, most of the reviews are scored 5 and evey scores else is really small, this can cause the traning and evaluating process become untrustworthy. But we can boost the performance of model by creating more data for score from 1 to 4 to make it more balance.



Unfortunatly, it will cost me more time becase I will need to convert English reviews to Vietnamese reviews.

c. Training model

- Model Choice: The selected model for sentiment analysis is PhoBERT, a transformer-based language model pre-trained specifically for Vietnamese text data. PhoBERT is designed and optimized for the nuances of the Vietnamese language, ensuring high performance in sentiment analysis tasks. By leveraging knowledge from extensive datasets during pre-training, PhoBERT is adept at adapting to the specific nuances and intricacies of sentiment analysis tasks. This tailored approach not only enhances performance but also facilitates faster convergence during training, ultimately leading to more efficient and effective sentiment analysis results.
    - Transformer:
        - Self-Attention Mechanism: The core component of the Transformer architecture is the self-attention mechanism. It allows the model to weigh the importance of different words in a sequence when encoding or decoding information. Each word in the input sequence attends to all other words, and the attention scores determine the relevance or importance of each word to the others. This mechanism enables the model to capture long-range dependencies in the input sequence more effectively.

        - Multi-Head Attention: To capture different types of information and learn diverse representations, the Transformer employs multi-head attention. It consists of multiple attention heads, each learning different attention patterns and representations of the input sequence. Multi-head attention allows the model to attend to different parts of the input sequence simultaneously, enhancing its ability to capture complex patterns and relationships.

        - Positional Encoding: Since the Transformer architecture does not inherently maintain the order of elements in a sequence, positional encoding is used to provide positional information to the model. Positional encoding embeddings are added to the input embeddings to encode the position of each word in the sequence. This enables the model to understand the sequential order of words and their relative positions in the input sequence.

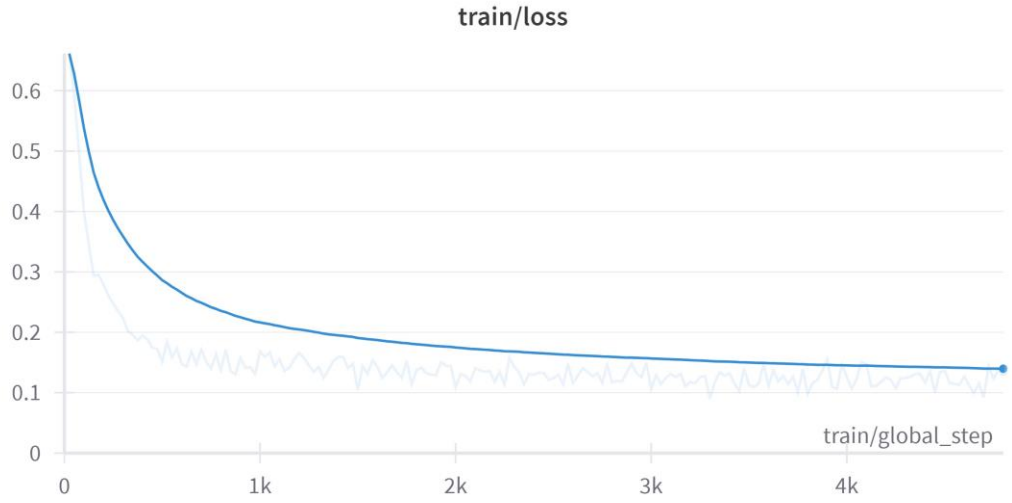
        - Feed-Forward Neural Networks: After applying self-attention and positional encoding, the Transformer architecture utilizes feed-forward neural networks to process the information. Feed-forward consist of multiple layers of fully connected networks with activation functions, allowing the model to learn non-linear transformations of the input features.

        - Layer Normalization and Residual Connections: To stabilize training and prevent the vanishing/exploding gradient problem, layer normalization and residual connections are applied after each sub-layer in the Transformer architecture. Layer normalization normalizes the activations across the feature dimension, and residual connections allow gradients to flow directly through the network, facilitating easier training of deeper architectures.

- Fine-tuning Approach: We wil continue fine-tuning Vietnamese Dataset of hotel reviews with PhoBERT. The config I am using for fine-tuning:
    - batch_size: 16
    - epoch: 5
    - learning_rate: 2e-5
    - warmup_steps: 500
    - weight_decay: 0.01
    - optimizer: AdamW

3. Presentation of model's performance including relevant metrics

- I am using 3 different evaluation metrics:
    - MSE: Measures the average squared difference between the predicted sentiment score and the true sentiment score, indicating the model's precision.
    - F1 Score: A weighted average of precision and recall, providing a balanced measure of model performance.
    - Accuracy: Measures the proportion of correctly predicted sentiment labels out of the total.
- My goal is to fine-tuning on 5 epochs, but after 2 epochs seem like the model stop decrease its lost then I stopped the training process.



- The evaluation of model after 1 epoch:
  {
  'eval_loss': 0.1342276632785797,
  'eval_mse': 25.241779327392578,
  'eval_accuracy': 0.8742334476665627,
  'eval_f1_score': 0.8563897639521286,
  'eval_runtime': 369.5222,
  'eval_samples_per_second': 26.036,
  'eval_steps_per_second': 1.629,
  'epoch': 1.0
  }

- After epoch 1, the training loss have not drop down, then it should be the same with the next epochs.